

---

# Passing Expectation Propagation Messages with Kernel Methods

---

Wittawat Jitkrittum    Arthur Gretton    Nicolas Heess \*

Gatsby Unit, University College London

wittawat@gatsby.ucl.ac.uk

arthur.gretton@gmail.com

nheess@gmail.com

## Abstract

We propose to learn a kernel-based message operator which takes as input all expectation propagation (EP) incoming messages to a factor node and produces an outgoing message. In ordinary EP, computing an outgoing message involves estimating a multivariate integral which may not have an analytic expression. Learning such an operator allows one to bypass the expensive computation of the integral during inference by directly mapping all incoming messages into an outgoing message. The operator can be learned from training data (examples of input and output messages) which allows automated inference to be made on any kind of factor that can be sampled.

## 1 Background

Existing approaches to automated probabilistic inference can be broadly divided into two categories (Heess et al., 2013): uninformed and informed cases. In the uninformed case, the modeler has full freedom in expressing a probabilistic model without any constraint on the set of usable factors. This high flexibility comes at a price during inference as less factor-specific information is available to the inference engine. Often MCMC-based sampling techniques are employed by treating the factors as black boxes. In the informed case (Stan Development Team, 2014; Minka et al., 2012), the modeler is required to build a model from constructs whose necessary computations are known to the inference engine. Although efficient during the inference, using an unsupported construct would require manual derivation and implementation of the relevant computations in the inference engine.

In this work, we focus on EP, a commonly used approximate inference method. Following Heess et al. (2013), we propose to learn a kernel-based message operator for EP to capture the relationship between incoming messages to a factor and outgoing messages. The operator bridges the gap between the uninformed and informed cases by automatically deriving the relevant computations for any custom factor that can be sampled. This hybrid approach gives the modeler as much flexibility as in the uninformed case while offering efficient message updates as in the informed case. This approach supports fast inference as no expensive KL divergence minimization needs to be done during inference as in ordinary EP. In addition, a learned operator for a factor is reusable in other models in which the factor appears. As will be seen, to send an outgoing message with the kernel-based operator, it is sufficient to generate a feature vector for incoming messages and multiply with a pre-learned matrix. Unlike Heess et al. (2013) which considers a neural network, the kernel-based message operator we propose can be easily extended to allow online updates of the operator during inference.

---

\*Nicolas Heess is now affiliated with Google Deepmind.

## 2 Expectation Propagation (EP)

Expectation propagation (Minka, 2001; Bishop, 2006) (EP) is a commonly used approximate inference method for inferring the posterior distribution of latent variables given observations. In a typical directed graphical model, the joint distribution of the data  $X = \{X_1, \dots, X_n\}$  and latent variables  $\theta = \{\theta_1, \dots, \theta_t\}$  takes the form of a product of factors,  $p(X, \theta) = \prod_{i=1}^m f_i(X|\theta)$  where each factor  $f_i$  may depend on only a subset of  $X$  and  $\theta$ . With  $X$  observed, EP approximates the posterior with  $q(\theta) \propto \prod_{i=1}^m m_{f_i \rightarrow \theta}(\theta)$  where  $m_{f_i \rightarrow \theta}$  is an approximate factor corresponding to  $f_i$  with the constraint that it has a chosen parametric form (e.g., Gaussian) in the exponential family (ExpFam). EP takes into account the fact that the final quantity of interest is the posterior  $q(\theta)$  which is given by the product of all approximate factors. In finding the  $i^{\text{th}}$  approximate factor  $m_{f_i \rightarrow \theta}$ , EP uses other approximate factors  $m_{\theta \rightarrow f_i}(\theta) := \prod_{j \neq i} m_{f_j \rightarrow \theta}(\theta)$  as a context to determine the plausible range

of  $\theta$ . EP iteratively refines  $m_{f_i \rightarrow \theta}$  for each  $i$  with  $m_{f_i \rightarrow \theta}(\theta) = \frac{\text{proj}[\int dX f(X|\theta) m_{X \rightarrow f_i}(X) m_{\theta \rightarrow f_i}(\theta)]}{m_{\theta \rightarrow f_i}(\theta)}$  where  $\text{proj}[r] = \arg \min_{q \in \text{ExpFam}} \text{KL}[r \| q]$  and  $m_{X \rightarrow f_i}(X) := \delta(X - X_0)$  if  $X$  is observed to be  $X_0$ . In the EP literature,  $m_{\theta \rightarrow f_i}$  is known as a cavity distribution.

The projection can be carried out by the following moment matching procedure. Assume an ExpFam distribution  $q(\theta|\eta) = h(\theta) \exp(\eta^\top u(\theta) - A(\eta))$  where  $u(\theta)$  is the sufficient statistic of  $q$ ,  $\eta$  is the natural parameter and  $A(\eta) = \log \int d\theta h(\theta) \exp(\eta^\top u(\theta))$  is the log-partition function. It can be shown that  $q^* = \text{proj}[r]$  satisfies  $\mathbb{E}_{q^*(\theta)}[u(\theta)] = \mathbb{E}_{r(\theta)}[u(\theta)]$ . That is, the projection of  $r$  onto ExpFam is given by  $q^* \in \text{ExpFam}$  that has the same moment parameters as the moments under  $r$ .

In general, under the approximation that each factor fully factorizes, an EP message from a factor  $f$  to a variable  $V$  takes the form

$$m_{f \rightarrow V}(v) = \frac{\text{proj}[\int d\mathcal{V} \setminus \{v\} f(\mathcal{V}) \prod_{V' \in \mathcal{V}} m_{V' \rightarrow f}(v')]}{m_{V \rightarrow f}(v)} := \frac{\text{proj}[r_{f \rightarrow V}(v)]}{m_{V \rightarrow f}(v)} := \frac{q_{f \rightarrow V}(v)}{m_{V \rightarrow f}(v)} \quad (1)$$

where  $\mathcal{V} = \mathcal{V}(f)$  is the set of variables connected to  $f$  in the factor graph. In the previous case of  $m_{f_i \rightarrow \theta}$ , we have  $\mathcal{V}(f) = \{X, \theta\}$  and  $V$  in Eq. 1 corresponds to  $\theta$ . Typically, when the factor  $f$  is complicated, the integral defining  $r_{f \rightarrow V}$  becomes intractable. Quadrature rules or other numerical integration techniques are often applied to approximate the integral.

## 3 Learning to Pass EP Messages

Our goal is to learn a message operator  $C_{f \rightarrow V'}$  with signature  $[m_{V \rightarrow f}]_{V \in \mathcal{V}(f)} \mapsto q_{f \rightarrow V'}$  which takes in all incoming messages  $\{m_{V \rightarrow f} \mid V \in \mathcal{V}(f)\}$  and outputs  $q_{f \rightarrow V'}(v')$  i.e., the numerator of Eq. 1. For inference, we require one such operator for each recipient variable  $V' \in \mathcal{V}(f)$  i.e., in total  $|\mathcal{V}(f)|$  operators need to be learned for  $f$ . Operator learning is cast as a distribution-to-distribution regression problem where the training set  $S_{V'} := \{([m_{V \rightarrow f}^n]_{V \in \mathcal{V}(f)}, q_{f \rightarrow V'}^n)\}_{n=1}^N$  containing  $N$  incoming-outgoing message pairs can be generated as in Heess et al. (2013) by importance sampling to compute the mean parameters  $\mathbb{E}_{r_{f \rightarrow V'}(v')} [u(v')]$  for moment matching. In principle, the importance sampling itself can be used in EP for computing outgoing messages (Barthelmé and Chopin, 2011). The scheme is, however, expensive as we need to draw a large number of samples for each outgoing message to be sent. In our case, the importance sampling is used for data set generation which is done offline before the actual inference.

The assumptions needed for the generation of a training set are as following. Firstly, we assume the factor  $f$  takes the form of a conditional distribution  $f(v_1|v_2, \dots, v_{|\mathcal{V}(f)|})$ . Secondly, given  $v_2, \dots, v_{|\mathcal{V}(f)|}$ ,  $v_1$  can be sampled from  $f(\cdot|v_2, \dots, v_{|\mathcal{V}(f)|})$ . The ability to evaluate  $f$  is not assumed. Finally we assume that a distribution on the natural parameters of all incoming messages  $\{m_{V \rightarrow f}\}_{V \in \mathcal{V}(f)}$  is available. The distribution is used solely to give a rough idea of incoming messages the learned operator will encounter during the actual EP inference. In practice, we only need to ensure that the distribution sufficiently covers the relevant region in the space of incoming messages.

In recent years, there have been a number of works on the regression task with distributional inputs, including Poczos et al. (2013); Szabo et al. (2014) which mainly focus on the non-parametric case and are operated under the assumption that the samples from the distributions are observed but not

the distributions themselves. In our case, the distributions (messages) are directly observed. Moreover, since the distributions are in ExpFam, they can be characterized by a finite-dimensional natural parameter vector or expected sufficient statistic. Hence, we can simplify our task to distribution-to-vector regression where the output vector contains a finite number of moments sufficient to characterize  $q_{f \rightarrow V'}$ . As regression input distributions are in ExpFam, one can also treat the task as vector-to-vector regression. However, seeing the inputs as distributions allows one to use kernels on distributions which are invariant to parametrization.

Once the training set  $S_{V'}$  is obtained, any distribution-to-vector regression function can be applied to learn a message operator  $C_{f \rightarrow V'}$ . Given incoming messages, the operator outputs  $q_{f \rightarrow V'}$  from which the outgoing EP message is given by  $m_{f \rightarrow V'} = q_{f \rightarrow V'} / m_{V' \rightarrow f}$  which can be computed analytically. We opt for kernel ridge regression (Schölkopf and Smola, 2002) as our message operator for its simplicity, its potential use in an online setting (i.e., incremental learning during inference), and rich supporting theory.

### 3.1 Kernel Ridge Regression

We consider here the problem of regressing smoothly from distribution-valued inputs to feature-valued outputs. We follow the regression framework of Micchelli and Pontil (2005), with convergence guarantees provided by Caponnetto and De Vito (2007). Under smoothness constraints, this regression can be interpreted as computing the conditional expectation of the output features given the inputs (Grunewalder et al., 2012).

Let  $\mathbf{X} = (\mathbf{x}_1 | \dots | \mathbf{x}_N)$  be the training regression inputs and  $\mathbf{Y} = (\mathbb{E}_{q_{f \rightarrow V'}^1} u(v') | \dots | \mathbb{E}_{q_{f \rightarrow V'}^N} u(v')) \in \mathbb{R}^{D_y \times N}$  be the regression outputs. The ridge regression in the primal form seeks  $\mathbf{W} \in \mathbb{R}^{D_y \times D}$  for the regression function  $g(\mathbf{x}) = \mathbf{W}\mathbf{x}$  which minimizes the squared-loss function  $J(\mathbf{W}) = \sum_{i=1}^N \|\mathbf{y}_i - \mathbf{W}\mathbf{x}_i\|_2^2 + \lambda \text{tr}(\mathbf{W}\mathbf{W}^\top)$  where  $\lambda$  is a regularization parameter and  $\text{tr}$  denotes a matrix trace. It is well known that the solution is given by  $\mathbf{W} = \mathbf{Y}\mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I})^{-1}$  which has an equivalent dual solution  $\mathbf{W} = \mathbf{Y}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top = \mathbf{A}\mathbf{X}^\top$ . The dual formulation allows one to regress from any type of input objects if a kernel can be defined. All the inputs enter to the regression function through the gram matrix  $\mathbf{K} \in \mathbb{R}^{N \times N}$  where  $(\mathbf{K})_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$  yielding the regression function of the form  $g(\mathbf{x}) = \sum_{i=1}^N a_i \kappa(\mathbf{x}_i, \mathbf{x})$  where  $\mathbf{A} := (a_1 | \dots | a_N)$ . The dual formulation therefore allows one to straightforwardly regress from incoming messages to vectors of mean parameters. Although this property is appealing, the training size  $N$  in our setting can be chosen to be arbitrarily large, making computation of  $g(\mathbf{x})$  expensive for a new unseen point  $\mathbf{x}$ . To eliminate the dependency on  $N$ , we propose to apply random Fourier features (Rahimi and Recht, 2007)  $\hat{\phi}(\mathbf{x}) \in \mathbb{R}^D$  for  $\mathbf{x} := [m_{V \rightarrow f}]_{V \in \mathcal{V}(f)}$  such that  $\kappa(\mathbf{x}, \mathbf{x}') \approx \hat{\phi}(\mathbf{x})^\top \hat{\phi}(\mathbf{x}')$  where  $D$  is the number of random features. The use of the random features allows us to go back to the primal form of which the regression function  $g(\hat{\phi}(\mathbf{x})) = \mathbf{W}\hat{\phi}(\mathbf{x})$  can be computed efficiently. In effect, computing an EP outgoing message requires nothing more than a multiplication of a matrix  $\mathbf{W}$  ( $D_y \times D$ ) with the  $D$ -dimensional feature vector generated from the incoming messages.

### 3.2 Kernels on Distributions

A number of kernels on distributions have been studied in the literature (Jebara and Kondor, 2003; Jebara et al., 2004). Relevant to us are kernels whose random features can be efficiently computed. Due to the space constraint, we only give a few examples here.

**Expected Product Kernel** Let  $\mu_{r^{(l)}} := \mathbb{E}_{r^{(l)}(a)} k(\cdot, a)$  be the mean embedding (Smola et al., 2007) of the distribution  $r^{(l)}$  into RKHS  $\mathcal{H}^{(l)}$  induced by the kernel  $k$ . Assume  $k = k_{\text{gauss}}$  (Gaussian kernel) and assume there are  $c$  incoming messages  $\mathbf{x} := (r^{(i)}(a^{(i)}))_{i=1}^c$  and  $\mathbf{y} := (s^{(i)}(b^{(i)}))_{i=1}^c$ . The expected product kernel  $\kappa_{\text{pro}}$  is defined as

$$\kappa_{\text{pro}}(\mathbf{x}, \mathbf{y}) := \left\langle \bigotimes_{l=1}^c \mu_{r^{(l)}}, \bigotimes_{l=1}^c \mu_{s^{(l)}} \right\rangle_{\otimes_l \mathcal{H}^{(l)}} = \prod_{l=1}^c \mathbb{E}_{r^{(l)}(a)} \mathbb{E}_{s^{(l)}(b)} k_{\text{gauss}}^{(l)}(a, b) \approx \hat{\phi}(\mathbf{x})^\top \hat{\phi}(\mathbf{y})$$

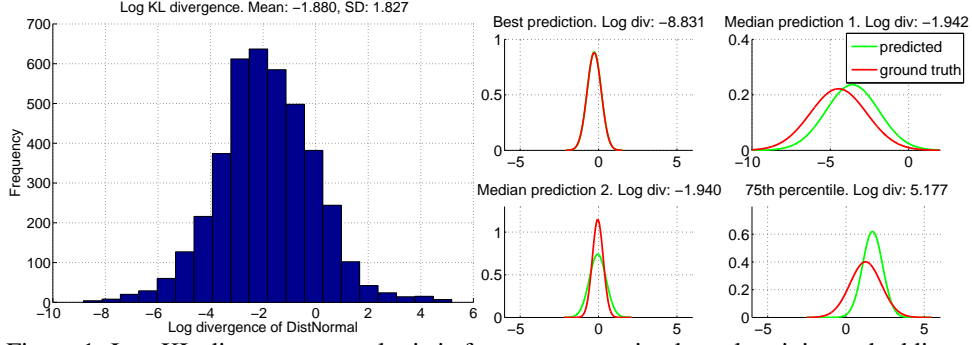


Figure 1: Log KL-divergence on a logistic factor test set using kernel on joint embeddings.

where  $\hat{\phi}(x)^\top \hat{\phi}(y) = \prod_{l=1}^c \hat{\phi}^{(l)}(r^{(l)})^\top \hat{\phi}^{(l)}(s^{(l)})$ . The feature map  $\hat{\phi}^{(l)}(r^{(l)})$  can be estimated by applying the random Fourier features to  $k_{\text{gauss}}^{(l)}$  and taking the expectations  $\mathbb{E}_{r^{(l)}(a)} \mathbb{E}_{s^{(l)}(b)}$ . The final feature map is  $\hat{\phi}(x) = \hat{\phi}^{(1)}(r^{(1)}) \otimes \hat{\phi}^{(2)}(r^{(2)}) \otimes \dots \otimes \hat{\phi}^{(c)}(r^{(c)}) \in \mathbb{R}^{d^c}$  where  $\otimes$  denotes a Kronecker product and we assume that  $\hat{\phi}^{(l)} \in \mathbb{R}^d$  for  $l \in \{1, \dots, c\}$ .

**Kernel on Joint Embeddings** Another way to define a kernel on  $x, y$  is to mean-embed both joint distributions  $r = \prod_{i=1}^c r^{(i)}$  and  $s = \prod_{i=1}^c s^{(i)}$  and define the kernel to be  $\kappa_{\text{joint}}(x, y) := \langle \mu_r, \mu_s \rangle_{\mathcal{G}}$  where  $\mathcal{G}$  is an RKHS consisting of functions  $g : \mathcal{X}^{(1)} \times \dots \times \mathcal{X}^{(c)} \rightarrow \mathbb{R}$  and  $\mathcal{X}^{(l)}$  denotes the domain of  $r^{(l)}$  and  $s^{(l)}$ . This kernel is equivalent to the expected product kernel on the joint distributions.

## 4 Experiment

As a preliminary experiment, we consider the logistic factor  $f(z|x) = \delta\left(z - \frac{1}{1+\exp(-x)}\right)$  which is deterministic when conditioning on  $x$ . The factor is used in many common models including binary logistic regression. The two incoming messages are  $m_{X \rightarrow f}(x) = \mathcal{N}(x; \mu, \sigma^2)$  and  $m_{Z \rightarrow f}(z) = \text{Beta}(z; \alpha, \beta)$ . We randomly generate 2000 training input messages and learn a message operator using the kernel on joint embeddings. Kernel parameters are chosen by cross validation and the number of random features  $D$  is set to 2000. We report  $\log KL[q||\hat{q}]$  where  $q = q_{f \rightarrow X}$  is the ground truth output message obtained by importance sampling and  $\hat{q}$  is the message output from the operator. For better numerical scaling, regression outputs are set to  $(\mathbb{E}_q[x], \log \mathbb{V}_q[x])$  instead of the expectations of the first two moments. The histogram of log KL errors is shown on the left of Fig. 1. The right figure shows sample output messages at different log KL errors. It can be seen that the operator is able to capture the relationship of incoming and outgoing messages. With higher training size, increased number of random features and well chosen kernel parameters, we expect to see a significant improvement in the operator’s accuracy.

## 5 Conclusion and Future Work

We propose to learn to send EP messages with kernel ridge regression by casting the KL minimization problem as a supervised learning problem. With random features, incoming messages to a learned operator are converted to a finite-dimensional vector. Computing an outgoing message amounts to computing the moment parameters by multiplying the vector with a matrix given by the solution of the primal ridge regression.

By virtue of the primal form, it is straightforward to derive an update equation for an online-active learning during EP inference: if the predictive variance (similar to a Gaussian process) on the current incoming messages is high, then we query the outgoing message from the importance sampler (oracle) and update the operator. Otherwise, the outgoing message is efficiently computed by the operator. Online learning of the operator lessens the need of the distribution on natural parameters of incoming messages used in training set generation. Determining an appropriate distribution was one of the unsolved problems in Heess et al. (2013).

## Acknowledgments

We gratefully acknowledge the support of the Gatsby Charitable Foundation.

## References

- S. Barthelmé and N. Chopin. Abc-ep: Expectation propagation for likelihood-free bayesian computation. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 289–296, New York, NY, USA, June 2011. ACM. ISBN 978-1-4503-0619-5.
- C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.
- A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, July 2007. ISSN 1615-3375. doi: 10.1007/s10208-006-0196-8. URL <http://dx.doi.org/10.1007/s10208-006-0196-8>.
- S. Grunewalder, G. Lever, A. Gretton, L. Baldassarre, S. Patterson, and M. Pontil. Conditional mean embeddings as regressors. In *ICML*, 2012.
- N. Heess, D. Tarlow, and J. Winn. Learning to pass expectation propagation messages. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3219–3227. 2013.
- T. Jebara and R. Kondor. Bhattacharyya and expected likelihood kernels. In *Conference on Learning Theory*. press, 2003.
- T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *J. Mach. Learn. Res.*, 5:819–844, Dec. 2004. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=1005332.1016786>.
- C. A. Micchelli and M. A. Pontil. On learning vector-valued functions. *Neural Comput.*, 17(1):177–204, Jan. 2005. ISSN 0899-7667. doi: 10.1162/0899766052530802. URL <http://dx.doi.org/10.1162/0899766052530802>.
- T. Minka, J. Winn, J. Guiver, and D. Knowles. Infer.NET 2.5, 2012. Microsoft Research Cambridge. <http://research.microsoft.com/infernet>.
- T. P. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, 2001. AAI0803033.
- B. Poczos, A. Rinaldo, A. Singh, and L. Wasserman. Distribution-free distribution regression. *arXiv:1302.0082 [cs, math, stat]*, Feb. 2013. URL <http://arxiv.org/abs/1302.0082>. arXiv: 1302.0082.
- A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Neural Information Processing Systems*, 2007.
- B. Schölkopf and A. J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, 2002.
- A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *In Algorithmic Learning Theory: 18th International Conference*, pages 13–31. Springer-Verlag, 2007.
- Stan Development Team. Stan: A c++ library for probability and sampling, version 2.4, 2014. URL <http://mc-stan.org/>.
- Z. Szabo, A. Gretton, B. Poczos, and B. Sriperumbudur. Consistent, two-stage sampled distribution regression via mean embedding. Feb. 2014. URL <http://arxiv-web3.library.cornell.edu/abs/1402.1754v2>.